Medical Record Linkage of Anonymous Registries without Validated Sample Linkage of the Dutch Perinatal Registries

Miranda Tromp, Nora Méray, Anita C. J. Ravelli, Johannes B. Reitsma, Gouke J. Bonsel

Abstract

This paper describes the linkage of data from three Dutch Perinatal Registries: the Dutch National Midwife Registry, the Dutch National Obstetrics Registry and the Dutch National Pediatrics Registry, for the year of 2001. All these registries are anonymous and lack a common identifier. We used probabilistic and deterministic record linkage techniques to combine data from the mother, delivery and child involving to the same pregnancy. Records of singleton and twin pregnancies were linked separately. We have developed a probabilistic close method based on maximum likelihood methods to estimate the weights of individual linking variables and the threshold value for the overall weight. Probabilistic linkage identified 80% more links than a full deterministic linkage approach. External validation revealed an error rate of less than 1%. Our method is a flexible and powerful method to link anonymous registries in the absence of a gold standard.

Keywords:
Medical Record Linkage; Probabilistic linkage; Medical registries; Perinatal care; Perinatal

1. Introduction

In the Netherlands four different caregivers: midwives, general practitioners, obstetricians and pediatricians are involved in perinatal care. All these four caregivers have their own anonymous registry: the registry of midwives (“MR”), of general practitioners (“GR”) of obstetricians (“OR”) and of pediatricians (“PR”); given the Dutch two-tier system. As in 2001 no information from the general practitioners was collected, only the MR, OR and PR registries were available for our study. About 30 to 40% of the pregnant women are treated by both midwives and obstetricians. 2% of the newborn children are admitted to the pediatrician. This means that information about the same mother/pregnancy/child often can be found in different perinatal registries. For studies focusing on outcomes of pregnancies (perinatal mortality and morbidity, etc.), it is essential that all information about mother, pregnancy, childbirth and child is available in one dataset. Linkage of the Dutch perinatal registries is therefore essential. Because of privacy laws in the Netherlands, no unique personal identifier exists that would enable linkage of records of different medical files.

Medical record linkage is the most commonly used technique to link records of the same individual in different files without a common identifier. In full deterministic MRL the value
of each linkage variables has to be identical in both records that are compared to be accepted as ‘link’ [1]. Deterministic linkage has a very high specificity, but low sensitivity. A refinement of deterministic linkage is the ‘n-1 deterministic linkage’, where one of the linkage variables is allowed to differ.

In probabilistic linkage a linkage weight is assigned to all independent linkage variables depending on the variable’s discriminating power and reliability [1-3]. This linkage weight is positive (reward) if the variable agrees in two compared records and has a negative value (penalty) in case of disagreement. The more discriminating power the variable has, the higher the linkage weight is in case of agreement. If a variable contains fewer errors, disagreement is less likely to occur among two related records, hence a higher negative weight (higher penalty). For every record pair the weight of the individual linkage variables is summarized in a total weight. This total weight represents the ratio of the probability of finding a particular pattern of agreements and disagreements among two records belonging to the same person to the same probability among two unrelated records. A “threshold value”, a total weight above which all records are accepted as links (and below which all pairs are considered as non links) has to be determined as well. The range around the threshold value is called ‘uncertain region’. Probabilistic compared to deterministic linkage trades a, hopefully small loss in with a higher gain in sensitivity.

This paper describes how probabilistic Medical Record Linkage (MRL) was applied to link the Dutch perinatal anonymous registries for the year of 2001. In absence of a golden standard we used additional information from original health care records to validate our approach.

2. Materials and Methods

Preparations prior to linkage procedure

Prior to linkage, the timeframe of the data files had to be defined. We chose for a birthday-based timeframe: all MR, OR and PR records where the child was (assumingly) born in 2001 were taken into account. In a second step the registries were cleaned from double entries by deterministic linkage. Prior to subsequent linkage procedures, the records of singletons and twins were separated in all three registries. Twins have the same mother and the same pregnancy information, often they are born on the same day and have the same gender, therefore linking records of twin pregnancies is complex. As a last step prior to linkage procedures, we reduced the number of pairs to compare by blocking either on birthday of the mother or on birthday of the child. This means that only pairs in which the two records agree on the blocking variable were compared. All records that were not linked with the first blocking variable were linked again with blocking on ZIP code of the mother to reduce the number of false negative links due to missing values and typing errors in the blocking variable.

Variable weights and the threshold value

In this study we further refined probabilistic linkage by allowing close values as well. This method is called the probabilistic close linkage. Close is where two records disagree on a variable but the difference is small and might be related to a typing error or a to a different way of measuring the same variable. In case of close agreement the linkage weight has still a positive value, but it’s always lower than in case of full agreement. The range of values for close agreement was defined based on the data. Likelihood methods were used to estimate the weights associated with agreement and disagreement for each variable, and to obtain an optimal value for the threshold value (more details in [8]). First, we estimated the so called
mi- and ui-values which are the probability of agreement among matches (mi) and among non-matches (ui) for each linkage variable had to be calculated:

\[ u_i = \Pr(\text{variable agrees} | \text{non matches}) \]

\[ m_i = \Pr(\text{variable agrees} | \text{matches}) \]

where \( i = 1, 2, \ldots, k; \) \( k \) being the number of variables.

An estimate of the ui-value was obtained by calculating the chance of agreement for the ith variable among all record pairs based on the marginal distribution of its values in the two files; a valid approach as long as the number of true matches is very small compared to the total number of pairs. Estimation of the mi values is more difficult as it requires the true state of (a large sample) pairs to be known. In the absence of a validated sample, we estimated the mi values by analyzing the observed patterns of agreements and disagreements among all pairs. If the outcomes of the comparisons are independent between variables, the total log likelihood can be written as:

\[
\sum_p n(\gamma^p) \left\{ \log (\pi \prod_{i=1}^{k} m_{fi}^{\gamma^p} m_{ci}^{\gamma^p} (1 - m_{fi} - m_{ci})^{1-\gamma^p} + (1-\pi) \prod_{i=1}^{k} u_{fi}^{\gamma^p} u_{ci}^{\gamma^p} (1 - u_{fi} - u_{ci})^{1-\gamma^p} ) \right\} \]  

Where \( m_{fi} \) is the probability of full agreement of the ith variable among matches, \( m_{ci} \) is the probability of close agreement among matches, \( u_{fi} \) is the probability of full agreement among non-matches, \( u_{ci} \) is the probability of close agreement among non matches, \( \pi \) is the proportion of true matches among all possible record combinations, \( n(\gamma^p) \) the number of record pairs with pattern \( \gamma^p \), \( \gamma^p_i \) is the outcome of the comparison of variable i in the pattern p, for \( i = 1, \ldots, k \) and \( p = 1, \ldots, 2^k \).

Using the mi- and ui-values that were estimated in equation 1, the variable weight of the “full agreement”, “close agreement” and “disagreement” can be calculated as:

- full agreement weight of the ith variable = \( \log_2 \frac{m_{fi}}{u_{fi}} \),
- close agreement weight of the ith variable = \( \log_2 \frac{m_{ci}}{u_{ci}} \),
- disagreement weight of the ith variable = \( \log_2 \frac{1 - m_{fi} - m_{ci}}{1 - u_{fi} - u_{ci}} \).

If a variable was missing in one or both of the records, the linkage weight for this variable was arbitrary set to 0 for this record pair.

**The linkage procedure**

The actual linkage procedure takes up to 13 different steps. Table 1 gives an overview of the linkage procedures. All steps of the record linkage were done with SAS software using standard procedures and self-written algorithms.

**External validation of MRL**

Two types of errors arise from MRL: false negative links where two records belonging to the same person are not linked; false positive links where two records not belonging to the same person are linked together [1]. Because no golden standard was available, we used an external validation procedure to determine the specificity and sensitivity of our linkage method. A stratified sample of the linked and non-linked pairs was chosen. We sought additional information from caregivers for each pair to determine the true status using information from medical records, discharge letter, personal communication, etc.
Table 1-13 independent linkage procedures were necessary to link the three registries of 2001

<table>
<thead>
<tr>
<th>Type of linkage procedure</th>
<th>Aim of linkage procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR with MR, deterministic linkage</td>
<td>Finding administrative double entries</td>
</tr>
<tr>
<td>OR with OR, deterministic linkage</td>
<td></td>
</tr>
<tr>
<td>PR with PR, deterministic linkage</td>
<td></td>
</tr>
<tr>
<td>PR singletons with PR singletons, probabilistic linkage</td>
<td>Linkage of different hospital entries of the same child</td>
</tr>
<tr>
<td>PR twins with PR twins, probabilistic linkage</td>
<td></td>
</tr>
<tr>
<td>MR singletons with OR singletons, probabilistic linkage</td>
<td>Linkage of midwife and obstetric records</td>
</tr>
<tr>
<td>MR twins with OR twins, probabilistic linkage</td>
<td></td>
</tr>
<tr>
<td>MR non-link singletons with OR non-link twins, prob. link.</td>
<td></td>
</tr>
<tr>
<td>MR non-link twins with OR non-link singletons, prob. link.</td>
<td></td>
</tr>
<tr>
<td>(MR^OR linked) singletons with PR singletons, prob. link.</td>
<td>Linkage of the linked midwife and obstetric records with the neonatal care records</td>
</tr>
<tr>
<td>(MR^OR linked) twins with PR twins, probabilistic linkage</td>
<td></td>
</tr>
<tr>
<td>(MR^OR linked) singletons with PR twins, prob. linkage</td>
<td></td>
</tr>
<tr>
<td>(MR^OR linked) twins with PR singletons, prob. linkage</td>
<td></td>
</tr>
</tbody>
</table>

3. Results

Due to space limitation we only present the results of the MR^OR linkage involving singleton pregnancies. Other linkage procedures produce results similar in nature. Table 2 gives an overview about the linkage variable weights arising from the MR^OR linkage procedure.

Table 2-Full agreement, close agreement and disagreement weight of the matching variables for the MR^OR linkage

<table>
<thead>
<tr>
<th>Matching variable</th>
<th>$m_i$-value</th>
<th>$u_i$-value</th>
<th>$\text{Linkage weight}$</th>
<th>agree</th>
<th>disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother’s date of birth</td>
<td>0.9608</td>
<td>0.000833</td>
<td>Blocking</td>
<td>10.17</td>
<td>-4.67</td>
</tr>
<tr>
<td>Mother’s postcode</td>
<td>0.9474</td>
<td>0.3217</td>
<td></td>
<td>1.56</td>
<td>-3.69</td>
</tr>
<tr>
<td>Number of pregnancies</td>
<td>0.9721</td>
<td>0.00267</td>
<td></td>
<td>8.35</td>
<td>-3.69</td>
</tr>
<tr>
<td>Child’s expected date of birth, fullmatch close (± 7 days)</td>
<td>0.9719</td>
<td>0.00211</td>
<td></td>
<td>8.85</td>
<td>-5.26</td>
</tr>
<tr>
<td>Time schedule of birth</td>
<td>0.9736</td>
<td>0.0433</td>
<td></td>
<td>4.49</td>
<td>-5.18</td>
</tr>
<tr>
<td>Child’s gender</td>
<td>0.9866</td>
<td>0.5003</td>
<td></td>
<td>0.98</td>
<td>-5.22</td>
</tr>
<tr>
<td>Child’s birth weight, fullmatch close (± 10 gram)</td>
<td>0.9308</td>
<td>0.00413</td>
<td></td>
<td>7.82</td>
<td>-4.48</td>
</tr>
</tbody>
</table>

The total weight of a record pair: $W = \sum w_i$, where $w_i$ is the (dis)agreement weight of the $i^{th}$ variable. The optimal threshold value for the MR^OR singleton linkage that minimised the overall error rate was estimated at 7.0. Therefore, every record pair with a total weight of more than 7 was classified as link.

Probabilistic close linkage of the MR and OR singleton records increased the number of links from 41,673 to 76,050 record pairs (Table 3).
To validate the results of our probabilistic close method, we obtained additional information from caregivers of a stratified sample of 524 record pairs. Because of lost As not all records could be found by the caregivers (we asked data from three years before), or because of non-response, finally 339 pairs could be analysed. 84 pairs were assumed ‘certain links’ (total weight above 12). 90 pairs were ‘certain non-links’ (total weight below 2). 165 pairs were from the uncertain region (total weight between 2 and 12), from which 69 pairs were right above the threshold value, ‘just linked pairs’, and 96 pairs were right below the threshold values, ‘just non-linked pairs’. We have found 19 false negative and 2 false positive links in the uncertain region. There was 1 false negative link in the certain non-link region, which was caused by a combination of missing values and typing errors in the records. All record pairs from the certain links region proved to be correctly linked. Figure 1 shows the number of correct (non)links and false positive-, false negative links against the total linkage weight of the record pairs.

Linkage of the MR and OR singleton records consists of comparison of more than 3 million record pairs, from which 3,080,264 pairs were found in the certain non-link region, 75,781 pairs in the certain links region (99.65 % of all linked pairs) and 1954 pairs in the uncertain region (1685 just non-links and 269 just links). Validation revealed 19+2=21 false links from the 165 validated pairs in the uncertain region, which means roughly (21/165) 13% false margin around the threshold value (about 254 pairs). Because only 0.35% of the linked pairs are in the uncertain region and only 0.055 % of the non-linked pairs, it means that our total linkage procedure has a less than 1 % margin of error.

Figure - External validation of the linked MR^OR file of a stratified sample of links with a total weight > 12 (certain links); a total weight < 2 (certain non-link); a weight between 2 and 12 (grey zone)

Linkage of the MR and OR singleton records consists of comparison of more than 3 million record pairs, from which 3,080,264 pairs were found in the certain non-link region, 75,781 pairs in the certain links region (99.65 % of all linked pairs) and 1954 pairs in the uncertain region (1685 just non-links and 269 just links). Validation revealed 19+2=21 false links from the 165 validated pairs in the uncertain region, which means roughly (21/165) 13% false margin around the threshold value (about 254 pairs). Because only 0.35% of the linked pairs are in the uncertain region and only 0.055 % of the non-linked pairs, it means that our total linkage procedure has a less than 1 % margin of error.
4. Discussion

Linkage of the PR to find subsequent hospital entries of the same child and linkage of the PR registry with the linked MR^OR registry was based on the same method as described above. The main difficulty of these linkage processes was that while the pregnant woman is the determinator of the MR and OR registries, the PR registry has the child as determinator.

Data cleaning proved to be essential prior to all linkage procedures. Blocking was unavoidable in all cases, though we are aware that it always introduces false negative links. Linkage of twin records remains difficult because of the very limited number of discriminating variables of the brothers and sisters. Longitudinal linkage to link different pregnancies of the same mother is at this moment impossible because of the limited maternal identifying variables in the recent registries. For this type of linkage only the mother’s birthday and her ZIP code could be used, but the later one often changes during the years due to moving.

5. Conclusion

We succeeded to link the anonymous Dutch perinatal registries in the absence of a golden standard. Probabilistic close linkage was the best linkage procedure. External validation proved more than 99% validity of the linked data. All information about pregnancy, child birth and postnatal period is now included in one record. The probabilistic method described here is useful tool for linking anonymous data from registries in the absence of a golden standard is present.

6. Acknowledgements

We gratefully acknowledge the support and funding of the SPRN (Foundation of the Dutch Perinatal Registry).

7. References


8. Address for correspondence

Nora Méray & Anita Ravelli
Academic Medical Center, University of Amsterdam, Department of Medical Informatics, J1B-115-1
P.O. Box 22700, 1100 DE Amsterdam, The Netherlands
email: n.meray@amc.uva.nl & a.c.ravelli@amc.uva.nl